Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 18 (2013) 2269 – 2277

Procedia Computer Science

2013 International Conference on Computational Science

Quantifying Uncertainty in Phylogenetic Studies of the Slavonic Languages

Diana Nurbakova[a]*, Sergey Rusakov[a], Vassil Alexandrov[b]

[a]Perm State National Research University, 15 Bukireva Street, Perm 614007, Russia
[b]*ICREA and Barcelona Supercomputing Center, C/ Jordi Girona, 29, Edifici Nexus II, E-08034 Barcelona, Spain*

Abstract

We describe the application of Bayesian methods to accommodate the uncertainty problem in phylogenetic reconstruction with an example of the Slavonic languages family. Comparative studies of languages have lots in common with evolutionary biology studies. Stable linguistic characters (e.g. word forms from the basic vocabulary, grammar characters) can be used to construct DNA-like sequences that the phylogenetic reconstruction methods can then be applied to. Linguistic data is known to be a subject of noise and error of different kinds causing conflicting signals and uncertainty within a phylogeny. Bayesian methods help to quantify the uncertainty. The comparison with the Damerau-Levenshtein distance-based tree is also given.

Keywords: uncertainty, phylogeny, language evolution, Slavonic languages, Bayesian inference

1. Introduction

Phylogenetic tree is a graph that reflects the evolution of species or other entities that have a common ancestor. It can also be used to visualize the genetic relationships between languages, the most remarkable and important phenomenon of the humanity that evolves through time as human race develops. Linguistic data is known to be complex, noisy and a subject of error. Being inference from data a phylogeny is not usually known with certainty. How successfully we can untangle the relationships between languages depends on the stability of the linguistic characters that are used like genes sequences and the methods that this data is analyzed with. There are two main approaches to phylogenetic reconstruction: distance-based methods that calculate the

* Corresponding author. Tel.: +7-909-727-5345.
E-mail address: d.nurbakova@gmail.com.

doi:10.1016/j.procs.2013.05.398

measure of mismatches between taxa and statistical methods that usually use optimality criterion and hypothesis testing.

Among numerous methods of phylogenetic reconstruction, Bayesian methods are an essential tool to handle the uncertainty and have been widely used in the last decade [1-6].

2. Bayesian inference of the Slavonic languages phylogenies

Bayesian methods are now used in the field of bioinformatics, genetics, applied linguistics, etc. One of the reasons is the complex and noisy data used in these fields. Being fully probabilistic, Bayesian methods can handle the uncertainty of data and extract the information from the data. Bayesian framework incorporates the prior information. The main shortcoming of Bayesian methods is its computational insensitivity. It is to point out that analytical solution of Bayesian inference is almost impossible and numerical methods are applied. However, the latter face high-dimensional integration problems. The most used techniques are based on Markov chain Monte Carlo (MCMC) algorithm that can be parallelized. [7,8]

Probability is considered as a direct measure of uncertainty. In Bayesian paradigm on the basis of prior probabilities and the likelihood of the data, Bayesian methods produce posterior probabilities. According to the Bayes' rule, the posterior probability that is the conditional distribution of the parameter θ, $p(\theta \mid X)$, given the data X, can be computed using the formula:

$$p(\theta|X) = \frac{p(\theta)p(X \mid \theta)}{p(X)} \tag{1}$$

- $p(\theta)$ – the prior probability (distribution) of the parameter that may be taken on the basis of theoretical or other considerations, or previous experiments;
- $p(X \mid \theta)$ – the likelihood of data;
- $p(X)$ – normalizing factor, the marginal probability of the data that is obtained by integrating over the prior distribution:

$$p(X) = \int_{\Theta} p(X \mid \theta)p(\theta)d\theta \tag{2}$$

- Θ – the parameter space for θ.

Therefore, the posterior probability can be represented as being proportional to the product of the prior distribution and the likelihood:

$$p(\theta \mid X) \propto p(\theta)p(X \mid \theta) \tag{3}$$

In application to phylogeny, Bayesian methods calculate a posterior probability of the node that reflects the proportion to the likelihood of trees in the sample with that node. [1, 2] In other words, given all observed data (cognates and morphological features), the hypothesis will be something like, "Ukrainian and Byelorussian are a group separate from Czech and Slovak", i.e. the former has a more recent common ancestor than any of them has with Czech or Slovak.

Given the observed taxa, the sample contains a set of the trees that occur most frequently. Thus, the posterior probability of tree T_i is calculated as follows:

$$p(T_i \mid \mathbf{M}) = \frac{p(\mathbf{M} \mid T_i)\, p(T_i)}{\sum_{\mathrm{T}} p(\mathbf{M} \mid T_i)\, p(T_i)} \tag{4}$$

- $p(T_i \mid \mathbf{M})$ is the conditional probability of tree T_i given $\boldsymbol{M}$;
- $\boldsymbol{M}$ is the sequence data represented in $n \times k$ matrix form, where n is the number of languages under consideration and k is the number of meanings (glosses) m_i. Each element of $\boldsymbol{M}$ is a numerical code to the word in n^{th} language that reflects the meaning. Any meaning could be described with cognate or non-cognate words. The former are assigned the same code. Thus, 'states' of a meaning are different non-cognate forms.
- $p(T_i)$ is the prior probability of tree T_i;
- $p(\mathbf{M} \mid T_i)$ is the likelihood of the data given the tree T_i (if meanings are independent of each other):

$$p(\mathbf{M} \mid T_i) = \int_{\mathbf{t}} \int_{\mathbf{Q}} p(\mathbf{t})\, p(\mathbf{Q})\, d\mathbf{t}\, d\mathbf{Q} \tag{5}$$

- $P(\mathbf{M} \mid \mathbf{Q}, T)$ is the probability (likelihood of the data) that the words from $\boldsymbol{M}$ evolve on a given phylogenetic tree:

$$P(\mathbf{M} \mid \mathbf{Q}, T) = \prod_i P(\mathbf{M}_i \mid \mathbf{Q}, T) \tag{6}$$

- $\boldsymbol{t}$ is the branch length;
- $p(\boldsymbol{t})$ is the prior probability of the branch length;
- $p(\boldsymbol{Q})$ is the prior probability of the parameter of the model;
- $\mathbf{Q} = \{q_{ij}\}_{(s+1)\times(s+1)}$ is the matrix of meaning changes that is written for any meaning with s states. q_{ij} reflects the instantaneous rate of change from state i to state j. The elements on the main diagonal are set $q_{ii} = -\sum_{\substack{j=0 \\ (j \neq i)}}^{s} q_{ij}$ so that the sum of elements in a row is always 0.0.

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ s \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & \cdots & s \end{array} \\ \begin{pmatrix} - & q_{01} & \cdots & q_{0s} \\ q_{10} & - & \cdots & q_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ q_{s0} & q_{s1} & \cdots & - \end{pmatrix} \end{array} \tag{7}$$

As far as a large number of combinations with finite data are possible, the uncertainty in topology and branch lengths should be taken into account [9].

As it was already mentioned, analytical solution may be very complicated or even impossible as calculation of the denominator $\sum_{\mathrm{T}} p(\mathbf{M} \mid T_i) p(T_i)$ is very computational intensive.

Markov Chain Monte Carlo (MCMC) methods are then used to generate a sample of trees with approximation of the posterior distribution of the trees reflecting the frequency sample distribution. A Markov chain is a random process that is characterized by transitions of a system from one state to another in a chainlike manner, where the next state depends only on the current one and not the priors (the Markov property). In phylogenetic paradigm, the states of a constructed Markov chain are different phylogenetic trees. In each iteration, a new tree is calculated by variation of branch lengths, tree topology, or the parameters of the model. Such trees are sampled using one of the sampling methods (e.g. Metropolis-Hastings algorithm, Gibbs sampling, slice sampling, and their different modifications). In MrBayes as well as in many others phylogenetic packages Metropolis-Hastings algorithm is implemented.

Metropolis-Hastings algorithm [10] is a sampling algorithm that uses an auxiliary distribution function $Q(x' \mid x^t)$, where x^t is the current state of the Markov chain. For this function it is easy to generate a sample. At each step, a random value x' for this function is generated. Then this new state is accepted as a next one $x^{t+1} = x'$ if $u > 1$, where:

$$u = \frac{P(x')Q(x^t \mid x')}{P(x^t)Q(x' \mid x^t)} \tag{8}$$

- $P(x)$ – any probability distribution from which samples are drawn.

Otherwise, the current state is retained: $x^{t+1} = x^t$.

Suppose, a random value of x is selected, x^t. In order to obtain the next value, a random value x' for the function $Q(x' \mid x^t)$ should be calculated first. Then the product $a = a_1 a_2$ is computed, where:

- $a_1 = \frac{P(x')}{P(x^t)}$ – the ration of probabilities of the interim value and the previous one,

- $a_2 = \frac{Q(x^t \mid x')}{Q(x' \mid x^t)}$ – the ration of probabilities of jumping from x' into x^t or retaining x^t. If Q is symmetric, then a_2 equals 1.

A random value on a new step is selected according to the following rule:

$$x^{t+1} = \begin{cases} x' & if\ a \geq 1 \\ x'\ with\ \ probability\ \ a & \\ x^t\ with\ \ probability\ \ (1-a) & if\ a < 1 \end{cases} \quad (9)$$

The algorithm starts from a random value x^0 and is run idle few steps in order to "forget" the initial value.

The best performance of the algorithm occurs in case if the form of the auxiliary function is close to the form of the target function.

MCMC implements a random walk. That determines the way a Markov chain is constructed: to have the integrand as its equilibrium distribution. Random walks are done between better and worse trees in the tree space. However, one of the properties of the chain that is the proportional visits of trees to their frequency of occurrence in the trees space provides a random trees sample. The frequency of the occurrence of a given monophyletic group in the MCMC trees sample is a Bayesian estimation of the posterior probability of actual existence of the node that defines the given group under consideration of the model of evolution and the given data.

3. Application to the Slavonic languages data set

The examined sample consists of the following Slavonic languages: Slovenian, Lower and Upper Lusatian, Czech, Slovak, Ukrainian, Byelorussian, Polish, Russian, Macedonian, Bulgarian and Serbo-Croatian.

According to the linguistic classification [11] the Slavonic languages are divided into three main branches due to their linguistic features and geographical distribution:

- East Slavonic: Russian, Byelorussian, Ukrainian;
- South Slavonic, which are further subdivided into:
  - Western subgroup: Serbo-Croatian and Slovene (or Slovenian);
  - Eastern subgroup: Bulgarian, Macedonian and Church Slavonic;
- West Slavonic, which are further subdivided into:
  - Lechitic languages: Polish, Kashubian, Silesian;
  - Sorbian (or Lusatian) languages: Upper and Lower Sorbian (Lusatian);
  - Czech and Slovak.

The close relationships between the Slavonic languages can be seen in their common synthetic language structure and in word structure, usage of the grammatical categories, syntactic structure, semantics, regular sound correspondences, morphological interchanges [12]. This closeness is explained by both, linguistic and extralinguistic, factors. The former implies the common origin and contacts of literary languages and dialects over a long period of time, while the latter deals with contacts of ethnic groups.

The most obvious difference between the West Slavic languages and the East ones is in orthography. Thus, the West Slavic languages that have been influenced mostly by Western Europe and their speakers being Roman Catholic use the Latin alphabet, whereas the East Slavic languages that have had more Greek and Byzantine influence and their speakers being Eastern Orthodox with Uniate faithful use the Cyrillic alphabet.

3.1. Lexical and morphological data

Here the genetic relationships between the Slavonic languages are examined using Swadesh's [13] basic vocabulary of 200 glosses. The word forms within a subset of the Kruskal et al.'s database [14] are classified according to their relationships into cognates (words derived from the common ancestor), doubtfully cognates and non-cognates.

Each meaning was then transformed to a sequence of binary characters that code cognate forms as “1” and non-cognate word forms as “0”. Thus, the data set of 476 characters was obtained for the synchronic data and the data set of 411 characters was obtained for the diachronic data. An example of the data is given below (see Table 1).

Table 1. An example of lexical data

Gloss	Czech	Russian	Ukrainian	Lower Lusatian	Slovak
all	vse	vse	uves'	wsen	vsetko
animal	zvire	zver	tvarina	zwerisco	zver
ashes	popel	pepel	popil	popel	popol

Morphological characters are considered stable (for example, see [15,16]). For a given language list of 12 languages, we picked five categories from WALS [17]. Unfortunately, the information stored in WALS is not the same for all languages and there are a lot of holes. Therefore, we added data from other linguistic sources to fill in the gaps in order to have less missing data. As the observing categories, Suffixing in inflexional morphology, Suppletion according to tense and aspect, Presence of definite articles, 3 genders, and Retention of the Dual number have been chosen representing different aspect of morphological data. This data was then coded with “1” in case of presence of the current character in a language and with “0” in case of its absence (see Table 2).

Table 2. Structural (morphological) data of 12 Slavonic languages. “?” is used to indicate missing data, “1” – to indicate the presence of a character, and “0” – its absence

	Strongly Suffixing	Suppletion according to tense and aspect	Definite articles	3 genders	Dual number
Slovenian	1	1	?	1	1
Lusatian_L	0	?	0	1	1
Lusatian_U	?	1	0	1	0
Czech	0	0	0	1	0
Slovak	?	1	?	1	0
Ukrainian	1	1	?	1	1
Byelorussian	1	1	0	1	0
Polish	1	1	0	1	0
Russian	1	1	0	1	1
Macedonian	0	0	1	1	0
Bulgarian	1	0	1	1	0
Serbocroatian	1	1	0	1	0

3.2. Bayesian tree of the Slavonic languages

As the Bayesian method implementation MrBayes software has been used [18]. In order to calculate a phylogenetic tree of Slavonic languages the method has been executed with the following parameters: given a

data set containing lexical and morphological characters for 12 languages. 1500000 generations of MCMC have been run.

The phylogenetic tree of the Slavonic languages obtained using Bayesian method is represented in Fig. 1. The relationships between languages could be traced in a clear way. Thus, Slovenian is rather far from the other languages and stays apart from all the other languages. Lower and Upper Lusatian make a very close group. They are connected with a pair of Czech and Slovak that are very close. One of the possible explanations of this closeness may be found in the common history of people that speak these languages as they have always been leaving neighbourly. Moreover, for 75 years they have been living in one state, Czechoslovakia and certainly, that has affected the languages.

Another branch is formed by Ukrainian and Byelorussian, and they stay close to Polish (the situation that have been discussed earlier). Russian had separated from these languages earlier and had its own way. However, these four languages make a separate branch that may be compared with the East Slavonic branch. Macedonian and Bulgarian make a close group that has a common ancestor with Serbo-Croatian (that might have separated before).

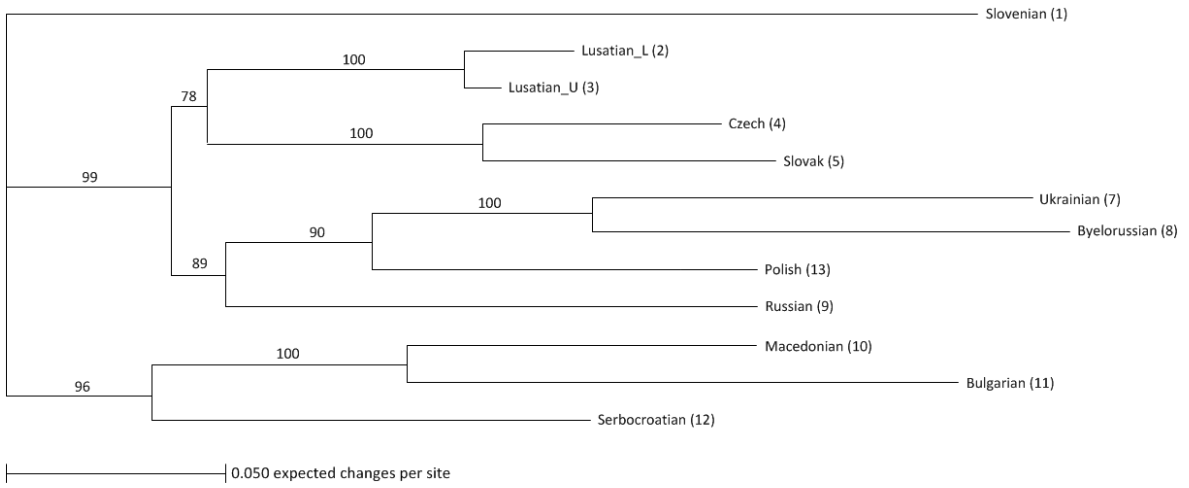

Fig. 1. MrBayes phylogenetic tree of 12 Slavonic languages on the basis of enlarged data set (lexical and morphological characters). Numbers above the internal branches show the posterior probability (%) of the nodes.

4. Comparison with the Damearau–Levenshtein distance-based tree

The Slavonic languages are known to be very close and that provokes a big quantity of cognates within the basic vocabulary. Though distance-based tree might be a subject of error giving false-positive or false-negative results caused by multiple phonetic changes and homonymous word forms, it seemed interesting to calculate the difference of the word forms that are very similar within the Slavonic languages. Different methods have been applied: maxmin algorithm, k-means, hierarchical clustering on the basis of correlation distance, Levenshtein and Damerau-Levenshtein distance as well as NeighbourNet algorithm.

Here we present the results obtained using the Damerau-Levenshtein distance [19,20] as it gave the most adequate results among the used methods and may rival Bayesian methods. To implement cluster analysis on the basis of Damerau-Levenshtein distance procedure in Wolfram Mathematica has been developed. Hierarchical clustering procedure has been applied to the dataset using Weighted Average linkage. The results are shown in Fig. 2.

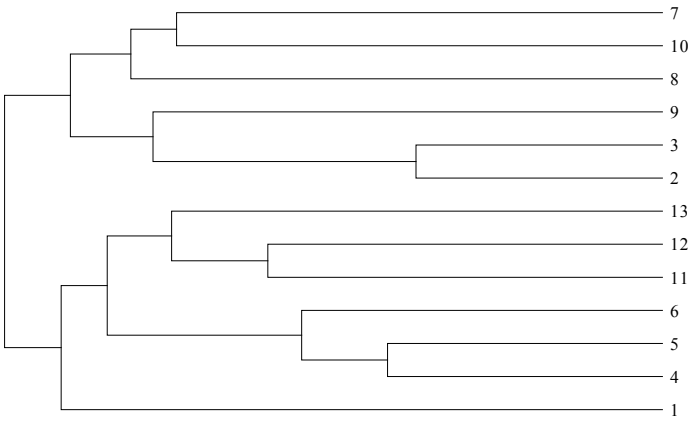

Fig. 2. The dendrogram of Slavonic languages on the basis of Damerau-Levenshtein distance. As the linkage technique the Weighted average is chosen. 1-Slovenian, 2-Lower Lusatian, 3-Upper Lusatian, 4-Czech, 5-Slovak, 6-East Czech, 7 - Ukrainian, 8-Byelorussian, 9 - Polish, 10-Russian, 11-Macedonian, 12-Bulgarian, 13-Serbo-Croatian

It may be noted that in this case Slovenian does not stay apart from all other Slavonic languages as in case of Bayesian tree but joins the South (Macedonian, Bulgarian and Serbo-Croatian) and West Slavonic languages (such as Czech and Slovak) which fits the linguistic classification of the Slavonic languages. According to the latter, Slovenian belongs to the South subgroup of the Slavonic languages but also has lots in common with the West Slavonic languages.

Lusatian languages are grouped together. And Polish joins them. This fact has an extralinguistic approval as Lusatian languages are spoken on the territory of Saxony, Brandenburg and the East-West part of Poland and are strongly influenced by the Polish language.

Ukrainian, Russian and Byelorussian show their relationships making the East Slavonic group. However, according to the tree, Ukrainian is closer to Russian than to Byelorussian which distinguishes that tree from the Bayesian one.

5. Conclusion

Our results obtained using Bayesian method does not contradict the classification given by scholars except ‘the migration’ of the Polish language from the West to the East Slavonic group. However this contradiction might reveal some deep relationships among the observed languages and might have an explanation in the history of these languages and peoples that speak these languages. Thus, Byelorussian and Ukrainian have undergone strong Polish influence as far as most part of the territory of modern Byelorussia and Ukraine belonged to the Polish-Lithuanian Commonwealth (or officially, Kingdom of Poland and Grand Duchy of Lithuania) and that had to affect languages a lot. Moreover, the Russian language that was spoken by Moscovy Kingdom had earlier separated from Byelorussian and Ukrainian when the capital of the state moved to Moscow.

This fact reminds us that any tree should be seen as a hypothesis of relationships between taxa as far as all models are simplified representations of real processes that are very complicated and complex.

References

[1] Hanson-Smith V, Kolaczkowski B, Thornton JW. Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty. *Mol. Biol. Evol.* 2010; **27**:1988-99.

[2] Pagel M, Lutzoni F. Accounting for phylogenetic uncertainty in comparative studies of evolution and adaptation. In Lässig M, Valleriani A, editors. *Biological Evolution and Statistical Physics*, Berlin, Springer-Verlag; 2002, p. 148-161.

[3] Huelsenbeck JP, Larget B, Miller RE, Ronquist F. Potential Applications and Pitfalls of Bayesian Inference of Phylogeny. *Syst. Biol.* 2002;**51**:673-688.

[4] Lartillot N, Poujol R. A Phylogenetic Model for Investigating Correlated Evolution of Substitution Rates and Continuos Phenotypic Characters. *Mol. Biol. Evol.* 2011; **28**:729-744.

[5] Nylander JAA, Olsson U, Alström P, Sanmartín I. Accounting for Phylogenetic Uncertainty in Biogeography: A Bayesian Approcah to Dispersal-Vicariance Analysis of the Thrushes (Aves: *Turdus*). *Syst.Biol.* 2008; **57**:257-268.

[6] de Villemereuil P, Wells JA, Edwards RD, Blomberg SP. Bayesian models for comparative analysis integrating phylogenetic uncertainty. *BMC Evolutionary Biology* 2012; **12**:102.

[7] Pagel M, Meade A. Bayesian estimation of correlated evolution across cultures: A case study of marriage systems amd wealth transfer at marriage. In Mace R, Holden CJ, Shannan S, editors. *The Evolution of Cultural Diversity: a phylogenetic approach*, London: University College London Press, 2005, p. 235-256.

[8] Holden CJ, Meade A, Pagel M. Comparison of Maximum Parsimony and Bayesian Bantu Language trees. . In Mace R, Holden CJ, Shannan S, editors. *The Evolution of Cultural Diversity: a phylogenetic approach*, London: University College London Press, 2005, p. 53-65.

[9] Gray RD, Atkinson QD. How old is Indo-European language family? Progree or more moth to the flame? In Renfrew C, editor. *Phylogenetic Methods and the Prehistory of Languages*, Cambridge: The McDonald Institute for Archaeological Research; 2006, p. 91-109.

[10] Greemberg E, Chib S. Understanding the Metropolis-Hastings Algorithm. *The American Statistician* 1995; **4**:327-335.

[11] Encyclopaedia Britannica, Slavic languages [http://www.britannica.com/EBchecked/topic/548460/Slavic-languages]

[12] Levitsky JA, Boronnikova NV. *History of Linguistics*. Moscow: Visshaya Shkola, 2005.

[13] Swadesh M. Lexicostatistic dating of prehistoric ethnic contacts. In *Proceedings of the American Philosophical Society* 1952: **96**:452-463.

[14] Kruskal JB, Black P, Dyen I. An Indo-European classification: a lexicostaatistical experiment. *Transactions of the American Philosophical Society* 1992; **82**.

[15] Warnow T, Evans S, Ringe D, Nakhleh L, Barbançon F., 2009. “An experimental study comparing lingiostic phylogenetic reconstruction methods. Languages and Genes,” Proceedings of the conference Languages and Genes, held at UC Santa Barbara.

[16] Warnow T, Ringe D, Evans SN, Nakhleh L. A Comparison of Phylogenetic Reconstruction Methods on an IE Dataset. *The Transactions of the Philological Society* 2005;**3**:171-192.

[17] Dryer MS, Haspelmath M editors. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library; 2011. Available online at http://wals.info/.

[18] Huelsenbeck J, Larget B, van der Mark P, Ronquist F, Simon D, Teslenko M. *MrBayes: Bayesian Inference of Phylogeny*. Available online at: http://mrbayes.sourceforge.net/.

[19] Levenshtein VI. Binary codes capable of correcting deletions, insertions and reversals. Soveit Scientists Doklady 1965;**163**:845-848.

[20] Chakrabarti S. *Mining the Web: Discovering Knowledge from Hypertest Data*. Morgan Kaufmann Publishers; 2003.